Animaltracker Data Validation: New Mexico Data

Joe Champion, Thea Sukianto

May 22, 2020

This document analyzes the results of the `animaltracker` package's data cleaning procedures by comparing data flagged by the app to data flagged by manual processing via spreadsheet.

The cleaning process uses flag-based rules for discarding cases (rows) of data.

- If measured rate of travel exceeds 84 m/min, mark the case with a `RateFlag`.
- If course change exceeds 100 degrees, mark the case with a `CourseFlag`.
- If measured distance traveled exceeds 840 m, mark the case with a `DistanceFlag`.
- Discard any case with a `DistanceFlag`, or 2+ flags (or both).

Preliminaries

Configure and load needed packages (use `install.packages("packagename")` to install any missing libraries).

```
library(dplyr)
library(ggplot2)
library(tidyr)
library(animaltracker)
library(psych)
```

Read and Prepare Data

```
###  read the manually cleaned data
clean_manual <- read.csv("df_correct.csv", stringsAsFactors = FALSE)

### read and clean the raw data with the animaltracker app
folder_rawdata <- "../test_data/DeepWell_2018_Collar_Raw"
nm_files <- list.files(folder_rawdata)

clean_anitracker <- data.frame() # container for cleaned data
for(filename in nm_files) {

  # extract metadata from file names
  aniid <- as.integer(gsub("DW_(\\d{3})(.*)", "\\1", filename))
  gpsid <- as.integer(gsub("DW_(\\d{3})_(\\d{2})(.*)", "\\2", filename))

  # read the raw data
  df_raw <- read.csv(file.path(folder_rawdata, filename), stringsAsFactors = FALSE)

  # clean with animaltracker
  df_clean_animaltracker <- clean_location_data(df_raw,
                                dtype = "igotu", filters = FALSE, maxtime =150,
                                aniid = aniid, gpsid = gpsid)

  # add to the combined clean data
  clean_anitracker <- rbind( clean_anitracker, df_clean_animaltracker)
```

```
}

### reshape data cleaned data to conform with manually cleaned data
clean_anitracker <- clean_anitracker %>%
  rename(Cow = Animal) %>% # use same name for cow id
  type.convert() # classify columns of data into types (e.g., numeric, factors)
```

First, we join the cleaned data from the animaltracker app (167901 rows, 34 columns) with the cleaned data from manual processing (167901 rows, 31 columns).

Rows are matched by the combination of `Cow`, `Index` (uniquely identifies almost all rows) and `Altitude` (to break ties in rare duplicates).

```
clean_anitracker <- clean_anitracker %>%
  arrange(Cow, Index, Altitude) %>%
  mutate(merge_index = 1:n())

clean_manual <- clean_manual %>%
  arrange(Cow, Index, Altitude) %>%
  mutate(merge_index = 1:n())

join <- full_join(clean_anitracker, clean_manual, by="merge_index") %>%
  rename(Index = Index.y,
                Cow = Cow.y,
                Altitude = Altitude.y,
                Order = Order.y,
                Keep.y = Keep,
                Speed = Speed.x,
                CourseDiff.x = CourseDiff,
                CourseDiff.y = coursedifference,
                DateTime = DateTime.x,
                Dist.x = Distance.x,
                Dist.y = Distance.y,
                DistFlag.x = DistanceFlag,
                DistFlag.y = DistFlag,
                MegaRateFlag.x = MegaRateFlag) %>%
  mutate( Cow = factor(Cow),
          Keep.x = 1*(TotalFlags.x < 2 & !DistFlag.x & !MegaRateFlag.x))
```

The merged data has 167901 rows.

Analysis

Overall Agreement

First, we compare the results of cleaning the data within `animaltracker` (via the `clean_location_data` function) to results of manual cleaning via spreadsheet.

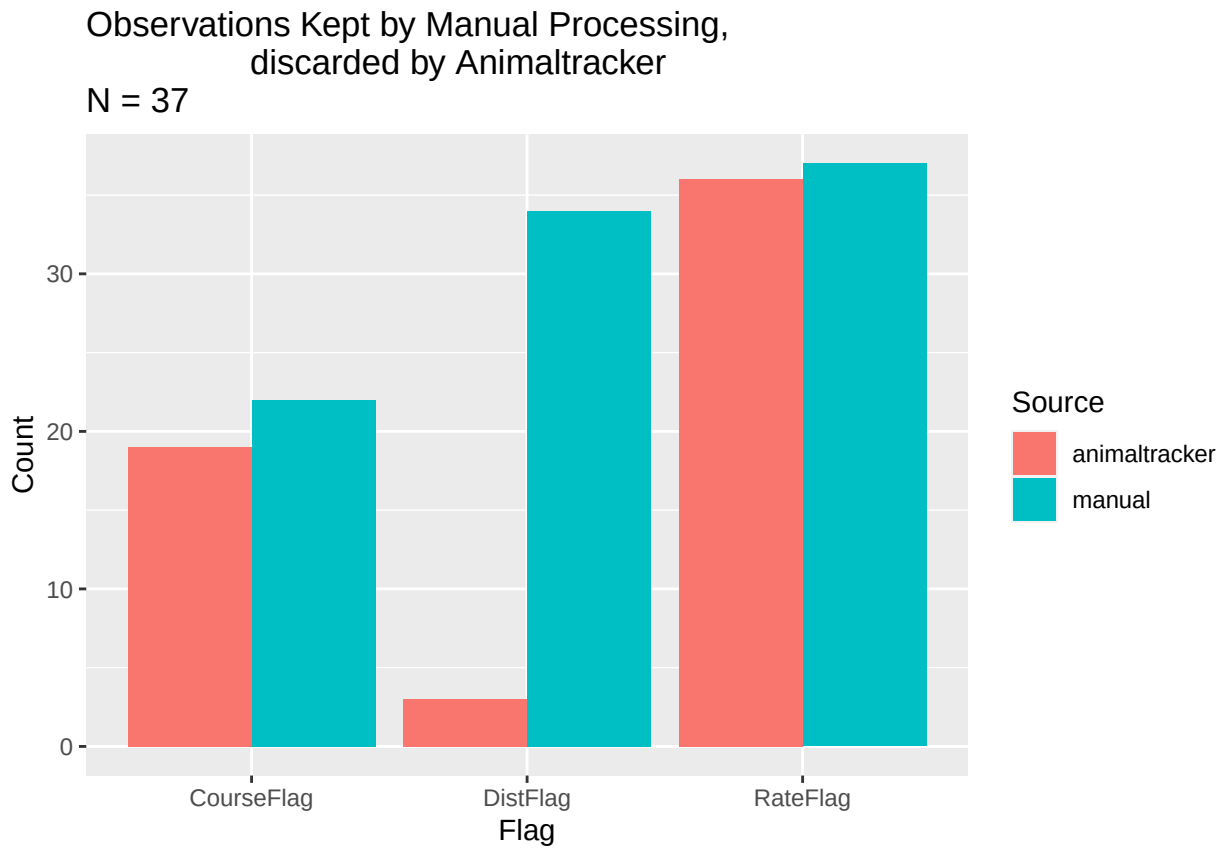
```
keepxtab <- with(join, table(Keep.x, Keep.y))
```

The cleaning methods agree in 99.84% of cases, except for 239 cases (0.14%) kept by `animaltracker` but discarded by manual processing and 37 cases (0.02%) kept by manual processing but discarded by `animaltracker`.

Analysis of Cases with Different Results

All cases kept by manual processing (n = 37) but discarded by `animaltracker` were marked with a `RateFlag` by manual, but not animaltracker.

```
manual_keep <- join %>%
  filter(Keep.x < Keep.y) %>%
  select(ind = merge_index, Cow, DateTime, TimeDiffMins,
              Rate.x, Rate.y, RateFlag.x, RateFlag.y,
              Dist.x, Dist.y, DistFlag.x, DistFlag.y,
              CourseDiff.x, CourseDiff.y, CourseFlag.x, CourseFlag.y)

manual_keep %>%
  summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  mutate(Source = ifelse(grepl(".x", Flag), "animaltracker", "manual"),
                Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot( aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle(paste0("Observations Kept by Manual Processing,
                 discarded by Animaltracker\n","N = ",nrow(manual_keep)) )
```



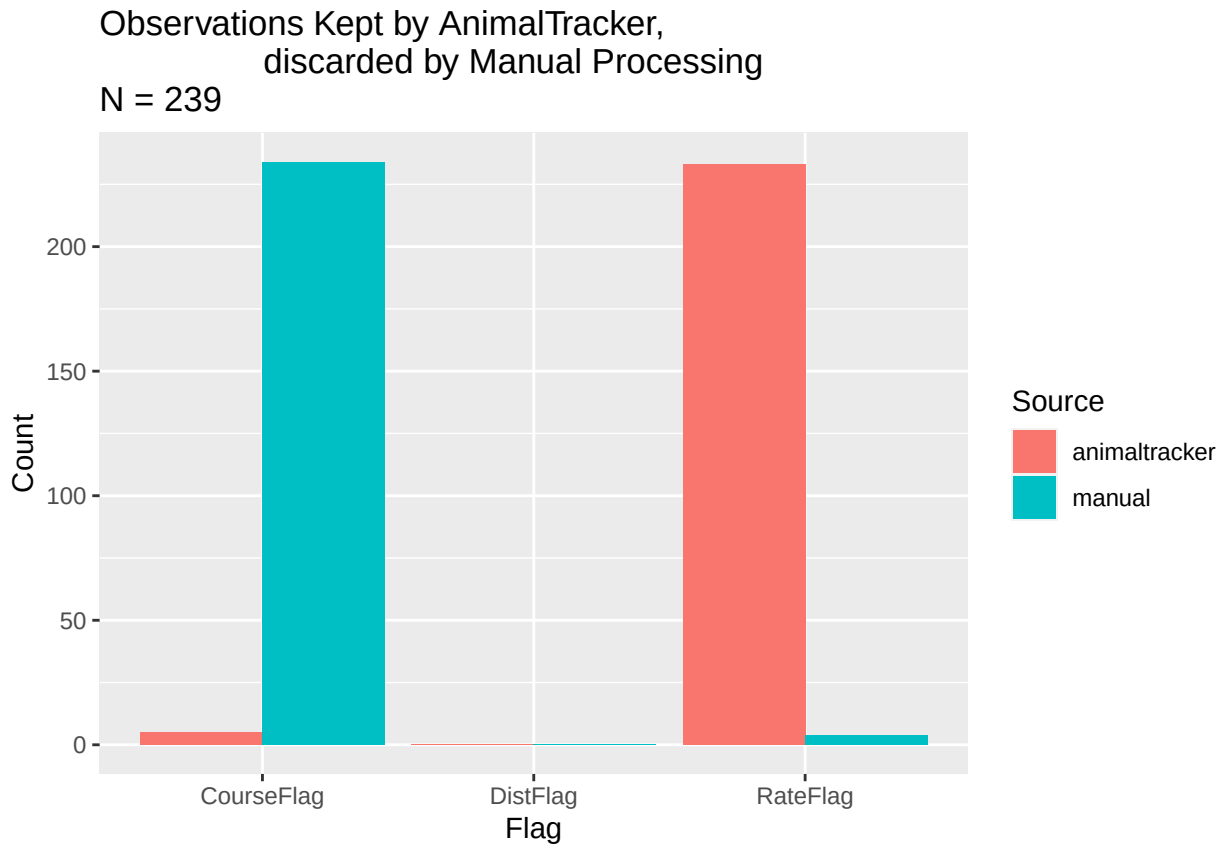
```
manual_keep %>% sample_n(10) # random sample of 10 cases
```

```
##      ind Cow            DateTime TimeDiffMins Rate.x      Rate.y
## 1  44037 225 2018-05-23 16:19:46            0    NaN           0
## 2  68270 229 2018-05-23 16:40:23            0    NaN     #DIV/0!
## 3  75650 257 2018-05-23 16:30:22            0    NaN           0
## 4  43500  63 2018-06-20 00:14:11            0    Inf 10.29568461
## 5  35669  63 2018-06-09 00:12:37            0    Inf 3.948926353
## 6  68269 229 2018-05-23 16:40:23            0    NaN     #DIV/0!
## 7  75617 229 2018-06-22 20:37:55            0    Inf        <NA>
## 8  75610 229 2018-06-22 19:47:40            0    NaN     #DIV/0!
## 9  41831  63 2018-06-17 16:01:48            0    Inf 12.84857532
## 10 68273 229 2018-06-12 18:06:58            0    Inf 9.327685888
##    RateFlag.x RateFlag.y   Dist.x   Dist.y DistFlag.x DistFlag.y
## 1           1          1  0.00000  0.00000          0          1
## 2           1          1  0.00000  0.00000          0          1
## 3           1          1  0.00000  0.00000          0          1
## 4           1          1 84.93940 84.93940          0          1
## 5           1          1 32.51283 32.51283          0          1
## 6           1          1  0.00000  0.00000          0          1
## 7           1          1 27.24322 27.24322          0          1
## 8           1          1  0.00000  0.00000          0          1
## 9           1          1 35.76187 35.76187          0          1
## 10          1          1 34.20152 34.20152          0          1
##    CourseDiff.x CourseDiff.y CourseFlag.x CourseFlag.y
## 1           217          217            1            0
## 2             0            0            0            1
## 3           184          184            1            0
## 4           141          141            1            0
## 5           224          224            1            0
## 6             0            0            0            1
## 7             0            0            0            1
## 8             0            0            0            1
## 9             0            0            0            1
## 10           29           29            0            1
```

Nearly all cases kept by `animaltracker` but discarded by manual processing (n = 239) had different values of `RateFlag` and `CourseFlag`.

```
anitracker_keep <- join %>%
  filter(Keep.x > Keep.y) %>%
  select(ind = merge_index, Cow, DateTime, TimeDiffMins,
              Rate.x, Rate.y, RateFlag.x, RateFlag.y,
              Dist.x, Dist.y, DistFlag.x, DistFlag.y,
              CourseDiff.x, CourseDiff.y, CourseFlag.x, CourseFlag.y)

anitracker_keep %>%
  summarise(RateFlag.x = sum(RateFlag.x),
                   CourseFlag.x = sum(CourseFlag.x),
                   DistFlag.x = sum(DistFlag.x),
                   RateFlag.y = sum(RateFlag.y),
                   CourseFlag.y = sum(CourseFlag.y),
                   DistFlag.y = sum(DistFlag.y)) %>%
  tidyr::gather("Flag", "Count") %>%
  mutate(Source = ifelse(grepl(".x", Flag), "animaltracker", "manual"),
```

```
              Flag = substr(Flag, 1, nchar(Flag)-2)) %>%
  ggplot( aes(Flag, Count, fill = Source)) +
  geom_bar(stat = "identity", position = "dodge") +
  ggtitle(paste0("Observations Kept by AnimalTracker,
                 discarded by Manual Processing\n","N = ",nrow(anitracker_keep)) )
```




```
anitracker_keep %>% sample_n(10) # random sample of 10 cases
```

```
##       ind Cow            DateTime TimeDiffMins    Rate.x      Rate.y
## 1   11308  11 2018-06-08 14:46:19     2.050000 102.79336 102.8078753
## 2  160803 535 2018-06-16 14:49:03     2.000000  86.68743 86.75985417
## 3   75539 229 2018-06-22 17:27:04     2.050000 164.96892 165.1632217
## 4    1575  11 2018-05-25 21:10:07     2.100000  88.68540 88.87416245
## 5   35002  63 2018-06-08 01:40:16     2.050000  85.56685 85.54504255
## 6   43879  63 2018-06-20 12:47:31     1.933333 100.68295 100.4613434
## 7   68182 225 2018-06-26 15:16:42     2.066667 383.36134  384.297384
## 8   17195  11 2018-06-16 23:43:50     2.000000 114.11892 113.9953566
## 9   74027 229 2018-06-20 15:14:36     1.983333 101.53146 101.6551537
## 10   1562  11 2018-05-25 20:43:02     2.083333  85.65587 85.59034731
##    RateFlag.x RateFlag.y   Dist.x   Dist.y DistFlag.x DistFlag.y
## 1           1          0 210.7561 210.7561          0          0
## 2           1          0 173.5197 173.5197          0          0
## 3           1          0 338.5846 338.5846          0          0
## 4           1          0 186.6357 186.6357          0          0
## 5           1          0 175.3673 175.3673          0          0
## 6           1          0 194.2253 194.2253          0          0
```

```
## 7           1          0 794.2146 794.2146          0          0
## 8           1          0 227.9907 227.9907          0          0
## 9           1          0 201.6161 201.6161          0          0
## 10          1          0 178.3132 178.3132          0          0
##    CourseDiff.x CourseDiff.y CourseFlag.x CourseFlag.y
## 1            14           14            0            1
## 2            13           13            0            1
## 3            14           14            0            1
## 4            13           13            0            1
## 5             1            1            0            1
## 6            21           21            0            1
## 7            13           13            0            1
## 8            45           45            0            1
## 9            14           14            0            1
## 10           31           31            0            1
```

Effects of Cleaning Differences on Outcome Measures

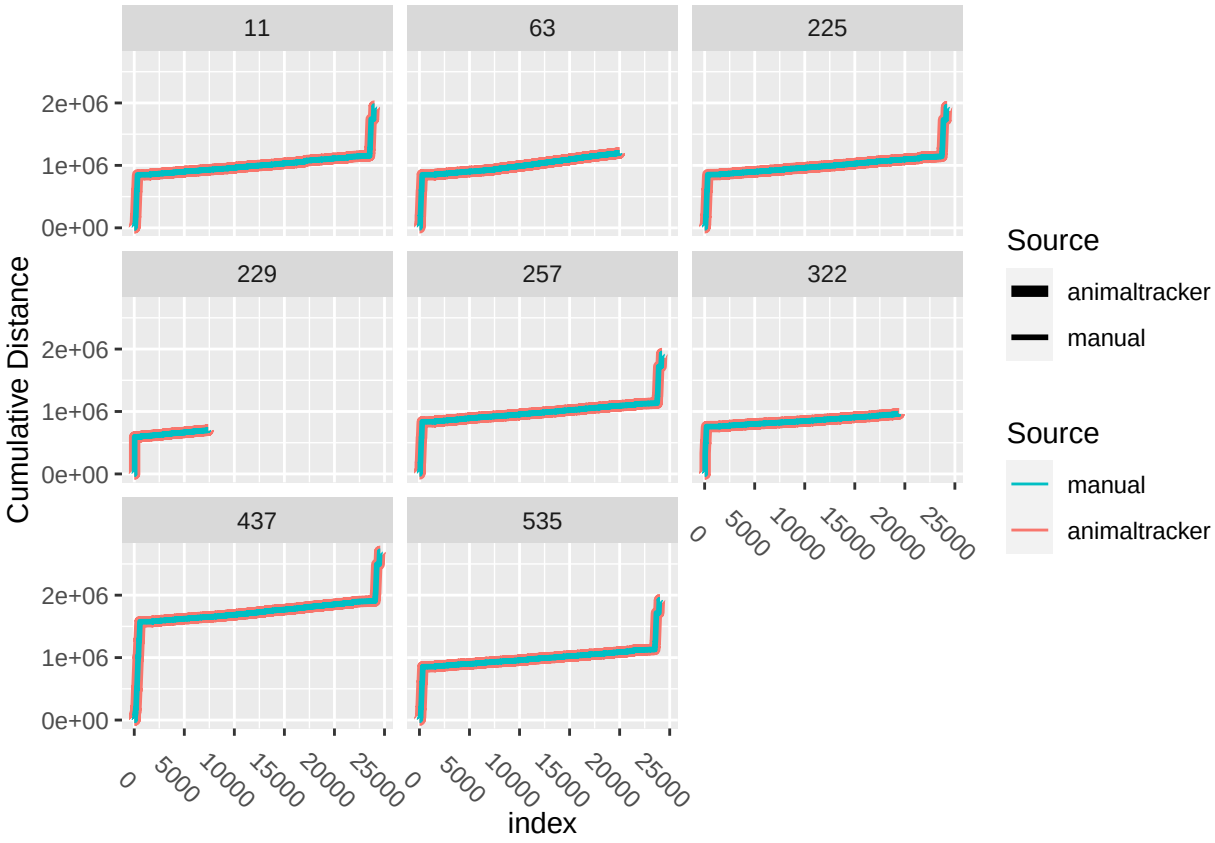
It's important to estimate the effects of processing errors on the key measured outcomes.

Cumulative Distance (per day)

The time series plots below indicate a very close conformity between the data cleaned in `animaltracker` and the manually cleaned data.

```
cumdist <- join %>%
  group_by(Cow) %>%
  arrange(merge_index) %>%
  mutate(Dist.y = lag(Dist.y,1),
                cumDist.x = cumsum(replace_na(Dist.x,0)),
                cumDist.y = cumsum(replace_na(Dist.y,0))) %>%
  ungroup()

cumdist_anitracker <- cumdist %>%
  group_by(Cow) %>% arrange(merge_index) %>%
  mutate(index = 1:n()) %>% ungroup() %>%
  ungroup() %>%
  select(Cow, index, cumDist.x, DistFlag.x) %>%
  rename(Flag = DistFlag.x,
                cumDist = cumDist.x) %>%
  mutate(Source = "animaltracker")

cumdist_manual <- cumdist %>%
  group_by(Cow) %>% arrange(merge_index) %>%
  mutate(index = 1:n()) %>% ungroup() %>%
  select(Cow, index, Cow, cumDist.y, DistFlag.y) %>%
  rename(Flag = DistFlag.y,
                cumDist = cumDist.y) %>%
  mutate(Source = "manual")

plot_data <- bind_rows(cumdist_anitracker, cumdist_manual)

ggplot(plot_data, aes(x=index, y=cumDist, group=Source, color=Source)) +
  geom_line(aes(size = Source)) +
  ylab("Cumulative Distance") +
```

```
  scale_color_discrete(guide = guide_legend(reverse = TRUE)) +
  scale_size_manual(values=c(2, 1)) +
  facet_wrap(vars(Cow)) +
  theme(axis.text.x = element_text(angle = -45))
```

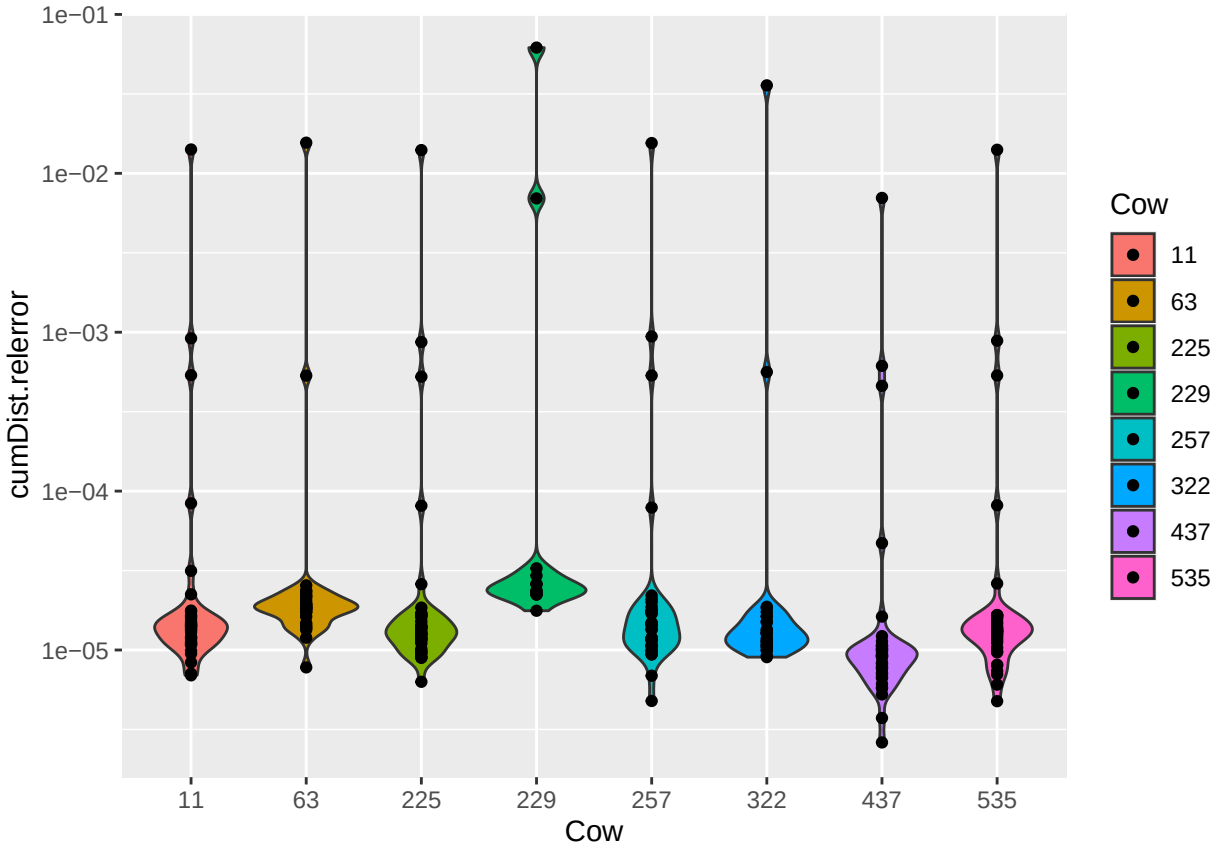


Relative Error of Cumulative Distance Estimates

The following summarizes the relative error of cumulative distances calculated from the `animaltracker` app in comparison to the manually processed data.

```
error_cumdist <- join %>%
  group_by(Cow) %>%
  arrange(merge_index) %>%
  mutate(
    Dist.y = lag(Dist.y,1),
    cumDist.x = cumsum(replace_na(Dist.x,0)),
    cumDist.y = cumsum(replace_na(Dist.y,0)) ) %>%
  group_by(Cow, Date.x) %>%
  summarize(
    cumDist.x = sum(cumDist.x, na.rm=TRUE),
    cumDist.y = sum(cumDist.y, na.rm = TRUE),
    cumDist.relerror = (cumDist.x-cumDist.y)/cumDist.y
  ) %>%
  ungroup()

ggplot(error_cumdist, aes(x = Cow, y = cumDist.relerror, fill = Cow))+
```

```
    geom_violin(trim=TRUE)+
    geom_point()+
    scale_y_continuous(trans='log10')
```



Based on N = 243 days of data, the overall relative error rate in cumulative distance per day is 0.08%.

```
error_cumdist %>%
  group_by(Cow) %>%
  mutate(index = 1:n()) %>%
  ungroup() %>%
  select(index, name = Cow, value = cumDist.relerror) %>%
  mutate(name = paste0("Cow_", name)) %>%
  pivot_wider() %>%
  select(-index) %>%
  psych::describe() %>%
  select(n, mean, sd, median, range, se ) %>%
  print(digits = 4)
```

```
##          n   mean     sd median  range     se
## Cow_11  35 0.0005 0.0024      0 0.0141 0.0004
## Cow_63  29 0.0006 0.0029      0 0.0156 0.0005
## Cow_225 35 0.0005 0.0024      0 0.0140 0.0004
## Cow_229 12 0.0058 0.0178      0 0.0618 0.0051
## Cow_257 35 0.0005 0.0026      0 0.0155 0.0004
## Cow_322 27 0.0014 0.0069      0 0.0358 0.0013
## Cow_437 35 0.0002 0.0012      0 0.0070 0.0002
## Cow_535 35 0.0005 0.0024      0 0.0141 0.0004
```

Rate of Travel

Another key outcome is the estimated speed of travel (meters/min). The following describes differences in estimated `Rate` measures between the data cleaned in `animaltracker` and the manually cleaned data.

```
rates_keep <- join %>%
    filter(Keep.x > 0 ) %>%
    select(merge_index, Rate = Rate.x) %>%
    mutate(source = "animaltracker") %>%
  rbind(
    join %>%
      filter(Keep.y > 0) %>%
      select(merge_index, Rate = Rate.y) %>%
      mutate(source = "manual")
  ) %>%
  mutate(source = factor(source),
         Rate = as.numeric(Rate))
```

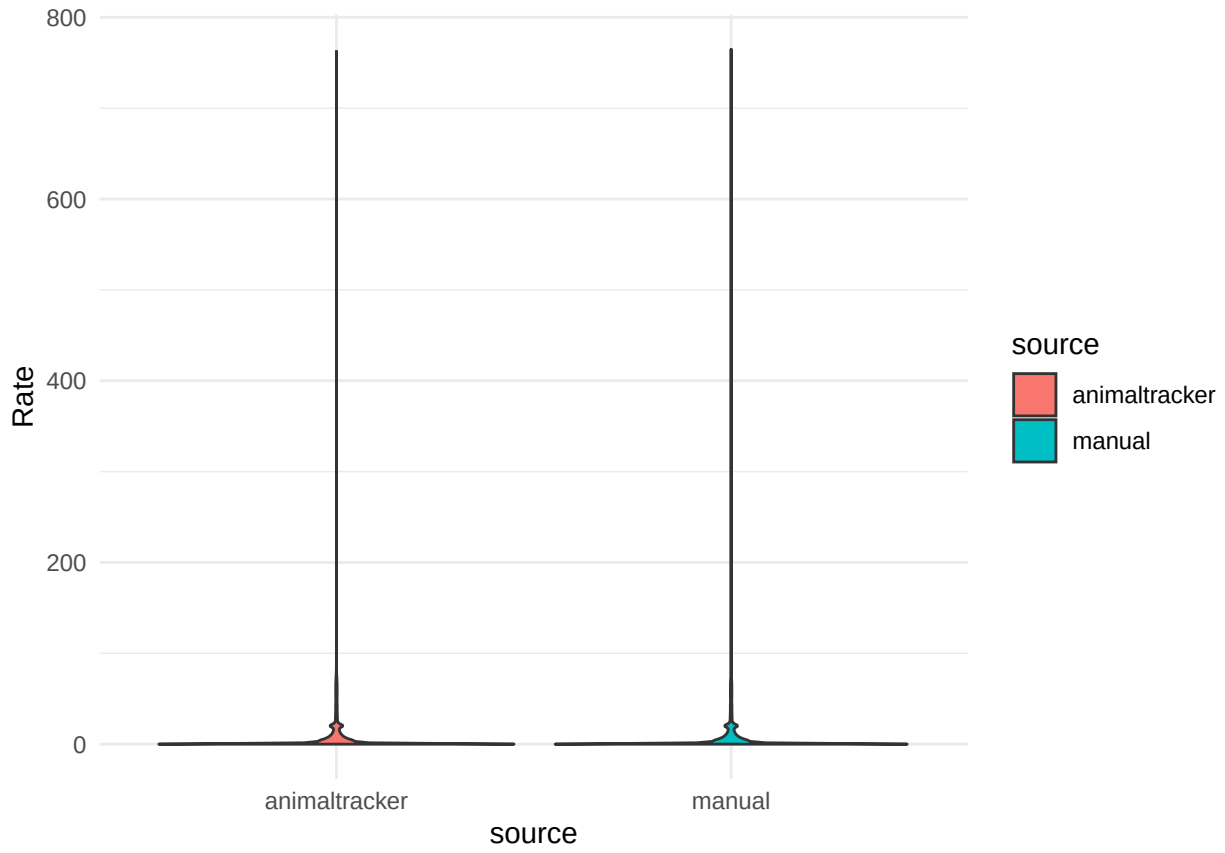
```
## Warning: NAs introduced by coercion
```

```
rates_keep %>%
  pivot_wider(names_from = "source", values_from="Rate") %>%
  select(-merge_index) %>%
  psych::describe() %>%
  select(n, mean, sd, median, range, se ) %>%
  print(digits = 3)
```

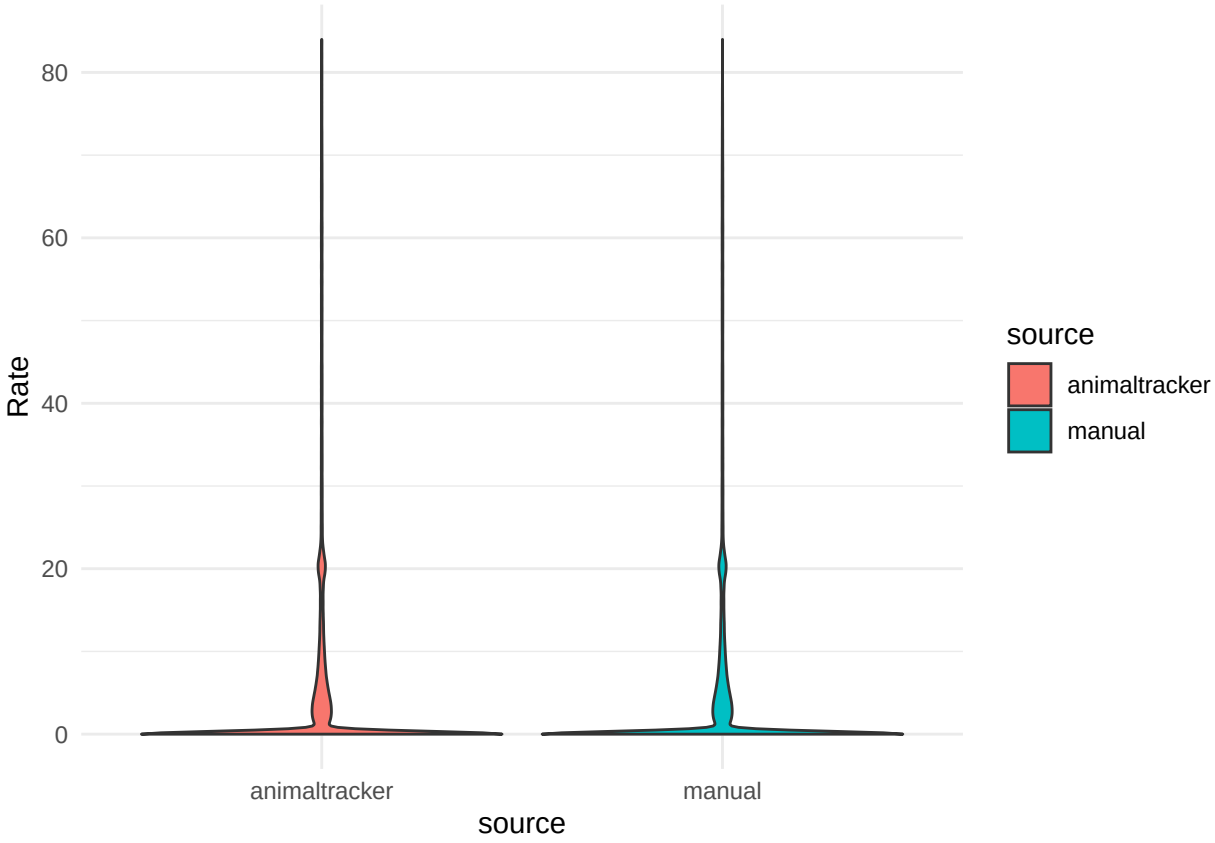
```
##                    n  mean     sd median   range    se
## animaltracker 165165 6.488 14.879      0 762.405 0.037
## manual        164970 6.297 13.696      0 764.775 0.034
```

```
ggplot(rates_keep, aes(x = source, y = Rate, fill = source))+
  geom_violin(trim=TRUE) +
  theme_minimal()
```

```
## Warning: Removed 18 rows containing non-finite values (stat_ydensity).
```




```
ggplot(rates_keep %>% filter(Rate < 84), aes(x = source, y = Rate, fill = source))+
  geom_violin(trim=TRUE) +
  theme_minimal()
```


80
60
40
20
0
Rate
animaltracker
manual
source
source
animaltracker
manual